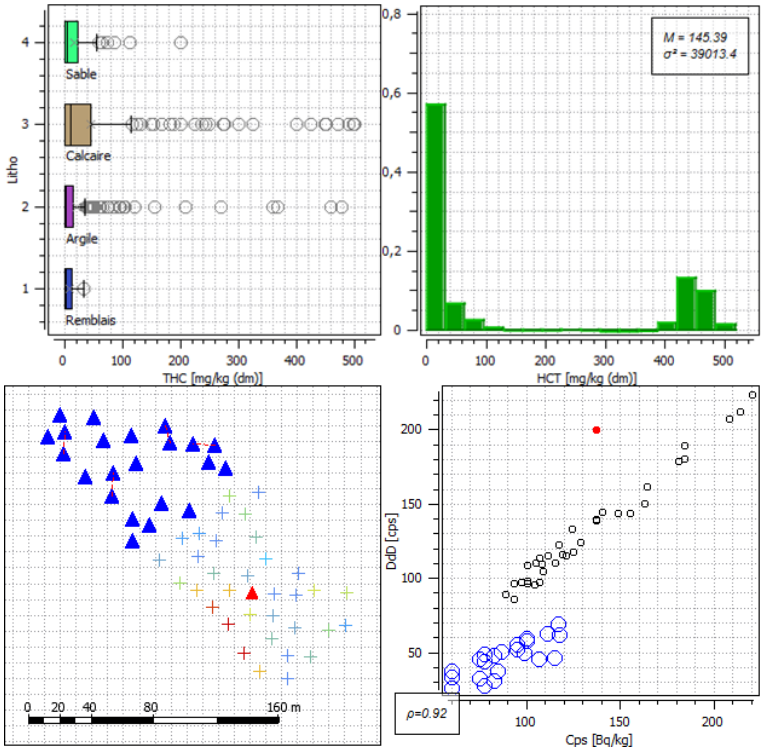
Geovariances
Where no one has gone before

Livre blanc



Contrôle qualité des données

Pourquoi est-il indispensable dans une étude de sites et sols pollués

Geovariances
Where no one has gone before

Contrôle qualité des données

Pourquoi est-il indispensable dans une étude de sites et sols pollués

Assurez-vous de la qualité de vos données en amont de l'étude à l'aide d'outils simples et visuels.

Avant d'entreprendre une évaluation de la contamination d'un site ayant pour objectif de fournir des outils d'aide à la décision pour sa réhabilitation, il est nécessaire de s'assurer de la qualité des données entrées. Ainsi, reposant sur des bases solides et fiables, les incertitudes au cours des différentes étapes du processus de décision seront mieux maîtrisées et la qualité finale du projet améliorée.

L'objectif de ce livre blanc est de montrer comment visualiser les données et de s'assurer au mieux de leur validité avant d'aller plus loin dans leur analyse.

Pour cela, différents outils peuvent être utilisés et sont abordés dans la suite : statistiques univariées, multivariées et outils de visualisation des données, qui sont ensuite avantageusement complétés par les statistiques spatiales et la géostatistique. Mis en œuvre par Geovariances depuis plus de quinze ans, ils ont montré leur caractère primordial dans ce type de projets.

Geovariances
Where no one has gone before

Le contrôle qualité des données : pourquoi ? comment ?

Êtes-vous confiant dans la fiabilité de vos données ?

La première étape d'une étude de sites et sols pollués est de vérifier la qualité des données afin d'assurer la fiabilité nécessaire au projet dans son ensemble. Ce contrôle intervient donc en amont d'une éventuelle modélisation afin de circonscrire au mieux les incertitudes dans le but de diminuer les aléas en cours de chantier et donc d'estimer le mieux possible les coûts du projet en diminuant le risque de mauvaises découvertes.

Les quelques outils statistiques et géostatistiques présentés dans la suite permettent de montrer comment passer d'un tableau de données brutes à des données validées en détectant différents types d'erreurs ou anomalies :

- erreurs de mesure,
- erreurs de recopie,
- anomalies dans la distribution statistique,
- erreurs dues à la localisation des échantillons : problèmes de coordonnées, problèmes de très forte hétérogénéité entre données proches, etc.

Par ailleurs, ces vérifications et corrections sont mises en œuvre de façon à toujours conserver le tableau original pour pouvoir s'y référer.

En réalisant ce travail, un second objectif est atteint : **les données sont mieux comprises** grâce à différentes représentations, par nature plus visuelles et informatives que de simples tableaux.

L'échantillonnage

Si le contrôle qualité des données est incontournable, établir une stratégie d'échantillonnage, réfléchie en amont, adaptée l'est tout autant.

La cohérence entre objectif d'évaluation, analyse de données et schéma d'échantillonnage est primordiale pour la réussite du projet.

Les outils de visualisation

Avant d'utiliser une quelconque représentation statistique, la visualisation du tableau de données peut déjà permettre une meilleure compréhension des données :

- tri selon une variable,
- sélection de certaines valeurs,
- mise en forme conditionnelle avec des couleurs (Fig. 1) permettant d'identifier des anomalies.

Sondage	Prof min	Prof max	THC	BTEX	Lithologie
O9	1	2	24,7		Clay
O9	2,5	4	28,25	<0,01	Limestone
P11	0	0,3	<0,25		Backfill
P11	2	2,5	<0,25		Clay
P11	2,5	4	109	<0,01	Limestone
P11	4	5	<0,25		Sand
P12	0,3	1,5	7,5	0,1	Clay
P12	1,5	2,5	50		Clay
P12	2,5	5	490		Limestone
P14	0	0,3	1,5	0,03	Backfill
P14	2	3,5	37,5		Limestone
P14	3,5	5	450	3,08	Limestone
P15	0	0,3	1,5		Backfill
P15	1,2	2	95	1,95	Clay
P15	2	3	62,5		Clay
P15	3	4,9	451	9,77	Limestone
P7	0	0,3	<0,25		Backfill

Fig. 1 : Exemple de mise en forme d'un tableau de données

Comment visualiser ses données ?

Si les données ont été **géoréférencées**, une simple carte d'implantation sur laquelle les valeurs mesurées peuvent être représentée avec des tailles ou des couleurs différentes (Fig. 2) permet très rapidement de détecter des problèmes de valeurs ou de localisation des échantillons tout en offrant une meilleure compréhension de l'organisation de la contamination.

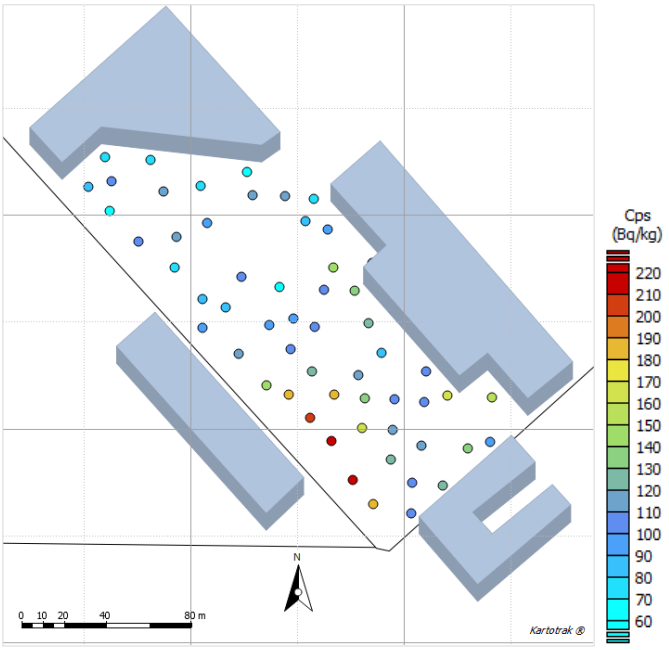


Fig. 2 : Carte d'implantation avec échelle de couleur en fonction de la variable d'intérêt radiologique

À trois dimensions, les logs de sondages peuvent venir compléter une carte d'implantation ou une vue 3D (Fig. 3), en montrant l'évolution de la contamination avec la profondeur, afin d'identifier les principaux niveaux impactés ainsi que la relation avec d'autres variables : autre polluant ou lithologie par exemple (Fig. 4).

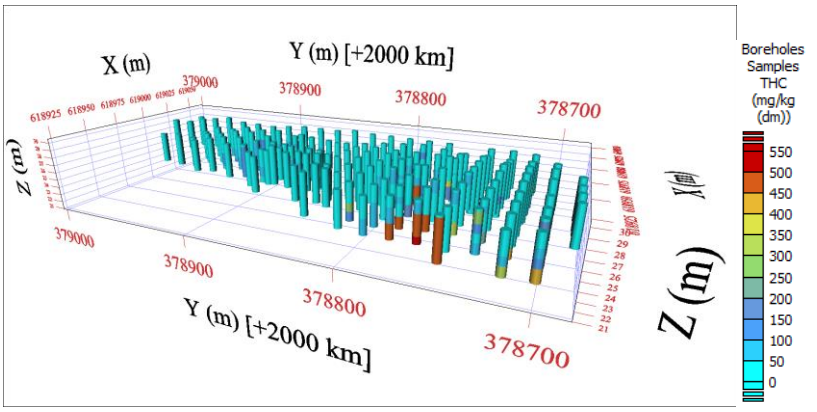


Fig. 3 : Visualisation à trois dimensions (hydrocarbures)

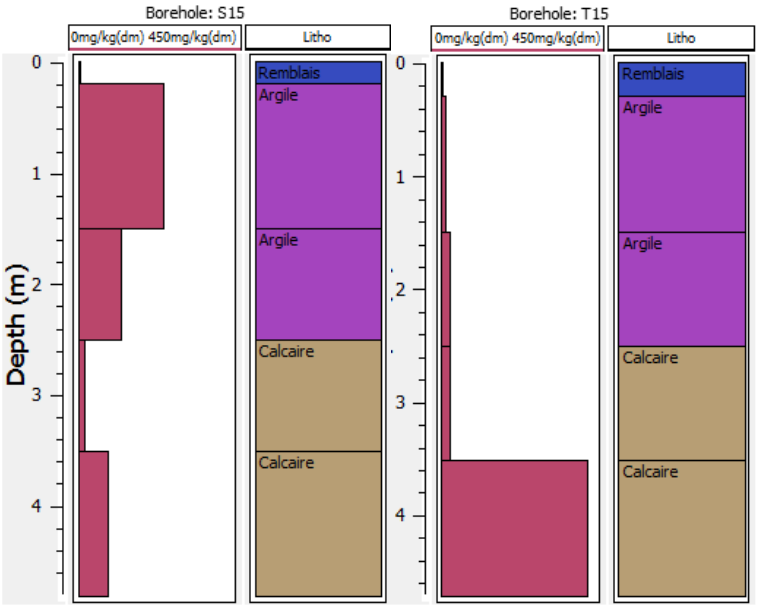


Fig. 4 : Logs de sondages de la variable d'intérêt et de la lithologie

Les outils statistiques

Comment détecter les valeurs aberrantes ?

Plusieurs outils statistiques permettent l'étude de la distribution des échantillons.

Parmi ceux-ci, **les histogrammes et les boîtes à moustaches** sont des outils permettant essentiellement d'étudier dans un premier temps une variable à la fois.

L'histogramme (Fig. 5) représente la distribution statistique des données. Dans les études de sites et sols pollués, il est courant d'observer la distribution de gauche ci-dessous : une forte proportion de valeurs faibles et quelques valeurs fortes. On peut

éventuellement y détecter des valeurs aberrantes ou des distributions inattendues.

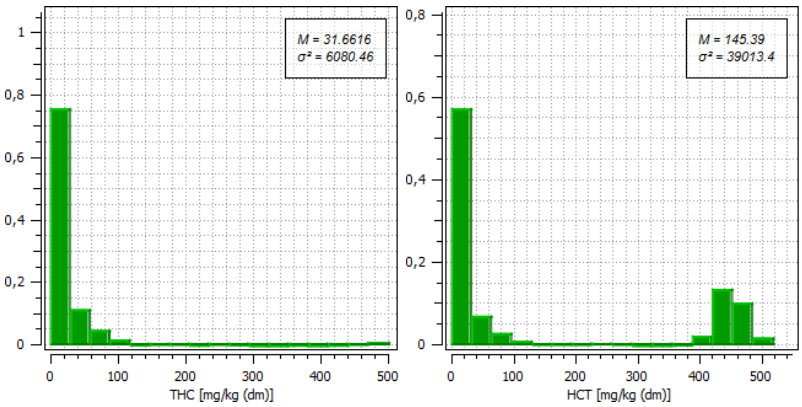


Fig. 5 : Exemples d'histogrammes (hydrocarbures)

L'histogramme de droite est dit bimodal. Il reflète en général l'existence de deux sous-populations. Un tel histogramme soulève des questions : s'attendait-on à cette distribution ? les échantillons ont-ils été tous analysés par le même laboratoire ? prélevés dans la même matrice ? avec le même support d'échantillonnage ? y a-t-il des différences de dates ?

Des effets de clustering ou d'échantillonnage préférentiel pourront également être détectés au moyen d'histogrammes, d'autant plus s'ils peuvent être dynamiquement reliés à une carte d'implantation (voir principe sur Fig. 7).

Les boîtes à moustaches peuvent venir compléter les histogrammes : plus synthétiques, ils permettent d'analyser la répartition des quantiles et surtout de voir les observations aberrantes. Réalisés par lithologie ou par domaine comme sur la Fig. 6, ils fournissent une information intéressante sur la répartition de la contamination, par exemple dans les différents horizons du sol.

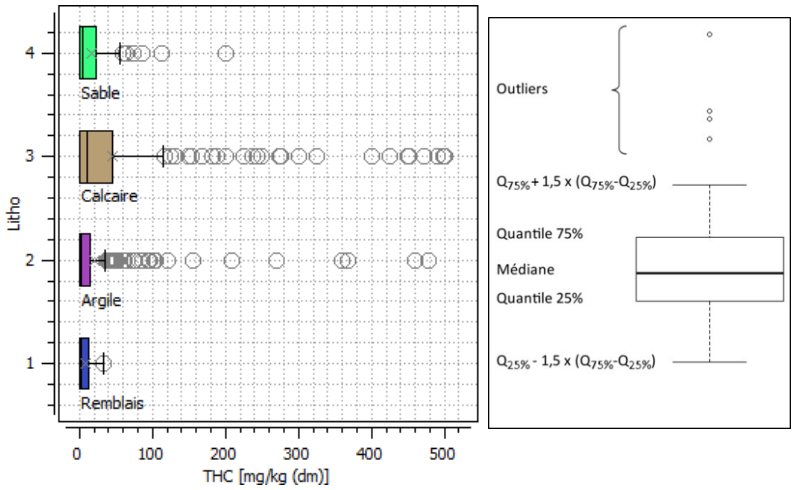


Fig. 6: Boîte à moustaches par lithologie (hydrocarbures) et définition de la boîte à moustaches

Certaines valeurs anormales peuvent ne pas être détectées à l'aide de ces outils univariés, par exemple dans les cas où certains échantillons présentent des valeurs anormales pour l'une ou l'autre variable. **Les nuages de corrélation** sont alors des outils intéressants pour étudier la relation entre les variables et détecter des valeurs anormales ou des nuages se décomposant en plusieurs parties symptomatiques de sous-populations.

La Fig. 7 représente la carte d'implantation et un nuage de corrélation entre deux variables. On y observe une relation linéaire avec une bonne corrélation. Cependant l'échantillon rouge sort de ce nuage, il semble présenter une valeur anormale pour l'une des deux variables (ce qui serait invisible sur un histogramme). On constate aussi un comportement différent dans le bas du nuage. Par sélection dynamique des points entre les deux graphiques, on observe que tous les points sont situés dans la même partie de l'aire étudiée. Ces deux points doivent soulever des questions et vérifications : la mesure du point rouge a-t-elle rencontré un problème ? s'agit-il vraiment d'un point très particulier ? y a-t-il une différence de laboratoire, de matrice, de date pour les deux sous-populations ? etc.

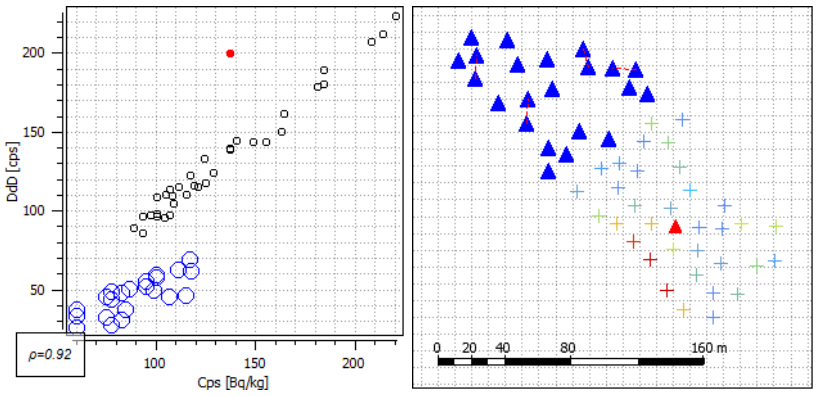


Fig. 7 : Nuage de corrélation (variables radiologiques) avec deux sous-populations et une valeur anormale (avec sélection correspondante sur la carte d'implantation)

Y a-t-il des incohérences dans la localisation des données ?

D'autres outils peuvent également être utilisés lors de cette phase préalable : analyse en composantes principales (ACP), classification… Plus poussés, ils sont néanmoins plutôt dédiés à des analyses statistiques avancées après validation des données.

Les outils : les statistiques spatiales et la géostatistique

Venant compléter les outils statistiques, **les statistiques spatiales et la géostatistique** permettent d'aborder un autre aspect du contrôle qualité des données : supposant une certaine cohérence spatiale de la pollution, existe-t-il des échantillons présentant des valeurs anormales compte tenu de leur localisation ?

Pour étudier cela, des outils comme la **nuée variographique** (représentation du carré des écarts entre deux points en fonction de la distance qui les sépare) et le **variogramme** (construit à partir de la nuée précédente, il représente la continuité spatiale du phénomène) sont particulièrement indiqués. Des indices tels qu'un effet de pépite (discontinuité à l'origine du variogramme qui traduit des écarts importants entre des points proches) très important ou des points élevés sur la nuée variographique correspondant à des paires sur la carte et mettant en jeu toujours le même échantillon (Fig. 8) peuvent révéler une valeur particulière compte tenu des valeurs voisines. Il conviendra alors non seulement de vérifier la valeur mesurée mais surtout sa localisation : erreurs de coordonnées X et Y, inversion ou translation dans les coordonnées, erreur de profondeur à trois dimensions…

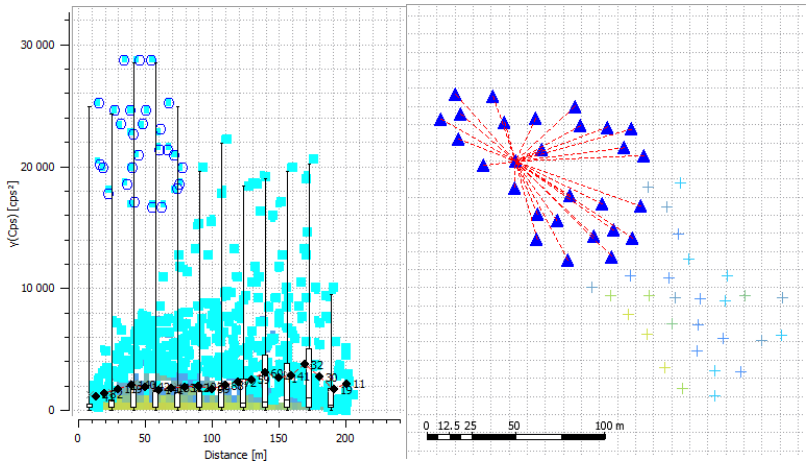


Fig. 8: Nuée variographique avec sélection de certains points correspondant à des paires impliquant un même échantillon sur la carte d'implantation

Enfin, le variogramme expérimental ne présente pas toujours la même continuité spatiale dans toutes les directions. En général, la variabilité spatiale est plus forte verticalement qu'horizontalement. Si une anisotropie horizontale est constatée elle est en général le reflet d'un phénomène physique tel qu'un écoulement. Si elle ne peut être expliquée, il sera nécessaire de bien vérifier que l'anisotropie n'est pas induite artificiellement par deux sous populations distinctes.

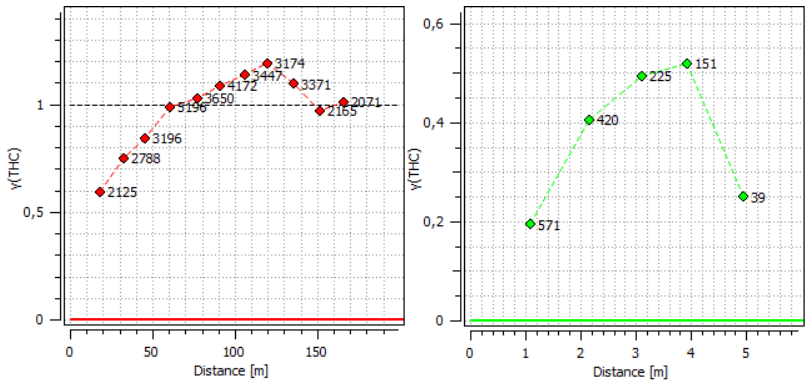


Fig. 9 : Variogramme horizontal en rouge et vertical en vert, reflet d'une anisotropie

Références

- Y. Desnoyers (Geovariances), D. Dubot (CEA) : Data Analysis for Radiological Characterisation: Geostatistical and Statistical Complementarity – Workshop on "Radiological Characterisation for Decommissioning", Studsvik, Sweden (April 2012)
- C. Faucheux, Y. Desnoyers (Geovariances) et P. De Moura (CEA) : Characterization of a deep radiological contamination: integration of geostatistical processing and historical data – AquaConsoil, Barcelona, Spain (April 2013)

Geovariances
Where no one has gone before

Qui est Geovariances ?

Geovariances combine une activité de développement logiciel, de bureau d'étude et de formation spécialisée en géostatistique. Nous avons plus de 45 employés, incluant des consultants en environnement et des statisticiens.

Geovariances développe et commercialise deux solutions logicielles :

- **Kartotrak** est une solution logicielle tout-en-un entièrement dédiée à la caractérisation de contaminations chimiques ou radiologiques.
- **Isatis** résulte de plus de 25 années d'expérience d'application industrielle de la géostatistique et constitue une solution logicielle complète pour toutes les questions géostatistiques.

Expertise unique

Geovariances est leader mondial dans le développement et l'application de solutions géostatistiques innovantes et pratiques. Nous avons une forte expérience dans la caractérisation de sites et avons gagné la confiance de leaders en environnement ainsi que de bureaux d'études.

Conclusion

Le contrôle qualité des données constitue la première étape indispensable d'une étude de sites et sols pollués. Pour le réaliser de nombreux outils simples et visuels existent.

L'histogramme et la boîte à moustaches permettent de détecter les valeurs anormales pour une variable. La représentation de nuages de corrélation peut faire apparaître des problèmes de valeurs compte tenu des relations existant entre les variables. Enfin les statistiques spatiales et la géostatistique viennent compléter ce contrôle en vérifiant la cohérence spatiale des valeurs.

Une fois réalisé, ce contrôle qualité assure une étude plus fiable du site et une éventuelle modélisation, avec une meilleure connaissance des incertitudes et rend la prévision d'un budget et d'un chantier plus fiable.

Notre expertise

Kartotrak est la première solution logicielle tout-en-un dédiée à la caractérisation des sites et sols pollués. Il est né d'un partenariat de plus de 10 ans entre le Commissariat à l'Energie Atomique et aux énergies alternatives et Geovariances. Simple d'utilisation, le logiciel propose une chaîne de traitement intégrée guidant l'utilisateur à chaque étape de son projet, depuis le chargement et le contrôle qualité des données, et la cartographie de la contamination, jusqu'à l'estimation des volumes de terres contaminés et l'évaluation des incertitudes.

Geovariances offre une expertise unique basée sur près de vingt années d'expérience dans l'application de la géostatistique aux problèmes de caractérisation et de réhabilitation de sites. La plupart de ses projets ont été réalisés à la demande des principaux donneurs d'ordre, institutionnels, bureaux d'études et entreprises de travaux.

Pour plus d'information

Nous sommes à votre disposition pour vous faire comprendre la valeur ajoutée des statistiques et de la géostatistique dans le cadre de vos projets de caractérisation de sites et sols pollués ou d'assainissement.

Contactez nos consultants : consult-env@geovariances.com.

Geovariances
49 bis, av. Franklin Roosevelt
77215, Avon Cedex
France
T +33 1 60 74 90 90
F +33 1 64 22 87 28

Geovariances Pty Ltd
Suite 3, Desborough House
1161 Hay Street
WEST PERTH, WA 6005
Australia
T +61 8 9321 3877

www.geovariances.com